COLUMBIA LAW REVIEW
SIDEBAR

VOL. 110 MARCH 11, 2010 PAGES 12–23

FEATURE SELECTION METHODS FOR SOLVING THE REFERENCE CLASS PROBLEM: COMMENT ON EDWARD K. CHENG, "A PRACTICAL SOLUTION TO THE REFERENCE CLASS PROBLEM"

James Franklin[*]

Response to: Edward K. Cheng, A Practical Solution to the Reference Class Problem, 109 Colum. L. Rev. 2081 (2009).

The reference class problem is illustrated by what Artificial Intelligence researchers call the Nixon Diamond.[1] Quakers are usually pacifists and Nixon is a Quaker. Republicans are usually not pacifists and Nixon is a Republican. On the total evidence, that Nixon is a Quaker and a Republican, what should we believe about whether Nixon is a pacifist?

The problem arises very generally, whenever statistical evidence is applied to an individual case. The case, Nixon, is a member of two different "reference classes"—the class of Quakers and the class of Republicans—and in these classes the frequency of the attribute or feature to be predicted—i.e., being a pacifist—differs. How then can we decide how the statistical evidence bears on the case? Should we attempt to find a best, most relevant, reference class? Or should we attempt to combine evidence from the various reference classes of which the case is a member, and if so, how?

In "A Practical Solution to the Reference Class Problem,"[2] Edward K. Cheng usefully surveys the ways in which the problem arises in legal contexts. In *United States v. Shonubi*,[3] sentencing guidelines required an estimate of how much heroin Charles Shonubi, a Nigerian drug

* Professor, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia.

1. Raymond Reiter & Giovanni Criscuolo, On Interacting Defaults: Proceedings of the 7th International Joint Conference on Artificial Intelligence 270 (1981).

2. Edward K. Cheng, A Practical Solution to the Reference Class Problem, 109 Colum. L. Rev. 2081 (2009).

3. 895 F. Supp. 460 (E.D.N.Y. 1995), discussed in Peter Tillers, If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by Their Reference Classes, 4 L. Probability & Risk 33 (2005).

smuggler, had carried through New York's John F. Kennedy Airport (JFK) on seven previous trips during which he had been undetected. The estimate was based on the average amounts of heroin found on Nigerian drug smugglers caught at JFK airport in the time period. Why should that be used as the reference class relevant to the case, rather than, say, George Washington Bridge tollbooth collectors (Shonubi's day job)? Or, take a more typical case involving valuation: Valuing a house for sale involves estimating its price from the sale records for "similar" houses. No other house is exactly the same as the given one, so how widely or narrowly should one choose the reference class of "similar" houses, and on what criteria? Number of bathrooms? Age? Street number?

Statistical theory has been dealing with inference from quantitative data for a very long time. So it is reasonable to hope, as Cheng argues, that the discipline of statistics will have available methods applicable to the reference class problem.[4]

It is true that traditional statistical theory tends to avoid the problem, taking for granted in expositions that the reference class has been correctly identified before inference begins. The basic idea of statistical inference is to observe counts—i.e., frequencies, proportions—in some reference class and apply the result as an estimate for a new, similar case. For example, one might draw a line of best fit through data of age and heights of trees in order to apply the estimated relationship between age and height to trees not yet observed. In expounding such techniques it is normally assumed that one has an unproblematic set of measurements of an identifiable and reasonably homogeneous set of trees. It is normally left to the statistician's good sense to choose a data set relevant to the problem.

But the recent expansion of statistics into the "data mining" of huge "data warehouses"[5] has forced consideration of how to identify what is relevant in large and mainly uninformative heaps of data, typically not collected with the current problem in mind. Cheng argues that a practical solution to the problem, at least when opposing counsel have put forward differing clear proposals on what the reference class should be, lies in modern "model selection" methods which decide on the appropriate complexity of a model by a formula such as the Akaike Information Criterion.[6] This Essay argues that a simpler area of recent statistics, the theory of feature selection methods, is more relevant. Since they are more straightforward and do not require an understanding of the issues concerning model complexity, they are

4. Cheng, supra note 2, at 2095 ("[T]he reference class problem is . . . a subspecies of the model selection problem [and] model selection criteria . . . eliminate the reference class problem as it arises in legal contexts.").

5. See generally Daniel T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining (2005); George M. Marakas, Modern Data Warehousing, Mining and Visualization: Core Concepts (2003).

6. Cheng, supra note 2, at 2093–94.

explained first, in Part I of this Essay. Part II discusses model complexity and argues that Cheng's approach is workable, but that the statistical literature provides some equally credible alternative approaches.

I. FEATURE SELECTION METHODS

The methods most applicable to problems like those in the legal context, such as real estate valuation, fall under the heading of "feature selection," also known as "variable selection" or "attribute selection."[7] A database is organized into many rows (the cases) and columns (the fields, attributes, properties, or features of the cases), as in the following schematic example:

ID	Address	Age (yrs)	Bedrooms	Area (sq ft)	Suburb median sale previous 12 months	Air conditioning?	. . .
1	1129 South Pkwy, Owens	12	4	9000	$1.2M	Y	
2	52 Central Ave, Springfield	40	2	4703	$440,000	N	
3	1 Liberty Ave, Springfield	5	3	4550	$660,000	Y	
. . .							

Table 1: Sample Real Estate Database

In general, one should imagine many more cases (rows): hundreds for real estate but millions for many kinds of health and gene data and

7. See generally Avrim L. Blum & Pat Langley, Selection of Relevant Features and Examples in Machine Learning, 97 Artificial Intelligence 245 (1997) (discussing feature selection in machine learning); Isabelle Guyon & André Elisseeff, An Introduction to Variable and Feature Selection, 3 J. Machine Learning Res. 1157 (2003) (same); Mark A. Hall & Geoffrey Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, 15 IEEE Trans. on Knowledge & Data Engineering 1437 (2003) (same); Patricia E.N. Lutu & Andries P. Engelbrecht, A Decision Rule-Based Method for Feature Selection in Predictive Data Mining, 37 Expert Sys. with Applications 602 (2010) (same).

financial records, and possibly thousands of features (columns). In such large cases, the great majority of features are expected to be irrelevant to the task of prediction. For example, not every feature or measurement in a gene database will be helpful in predicting cancer, and most features of financial records will be irrelevant to determining creditworthiness. The more features in a database, the harder it is to evaluate each feature's relevance.

The aim of feature selection methods is to determine from large amounts of data which of the many properties or features of the individual cases are relevant to a given classification or prediction task. For example, of the many features of houses in a real estate database—house size, lot size, number of bathrooms, street number, age, zip code, etc.—which are relevant to predicting the price?

One major purpose of identifying relevant features is to prevent the computations from becoming unfeasible, since computing with all the data, most of which is irrelevant, would be impossible for databases of the size typically used. But identifying the relevant features is also an aim in itself, since it will enable researchers to understand the data and form hypotheses as to what features are driving the system. These features could then either be classified as a symptom or identified as a cause and possibly changed.[8]

For the present purpose, however, the main significance of feature selection is that it largely solves the reference class problem. Choosing the relevant features determines the appropriate reference class for a case: Ideally, it is the class of those items that share with it all the features that have been found relevant for the task. However, that ideal is not always attainable, as discussed below. There is an accepted definition of the relevance of a feature to an outcome—for example, of "number of bathrooms" to "house price." A feature is relevant if it gives some information about the outcome—i.e., if "number of bathrooms" makes some difference to "house price" in the sense that, on average, a different number of bathrooms goes with a different house price. Relevance is correlation. In a very simple example, the reason that traffic lights are informative is that green is very highly correlated with it being safe to drive through the intersection: green and it is safe, red and it is not.[9] The color of the car ahead, however, is not correlated with the safe time to drive, so there is no point attending to it when deciding whether to drive across the intersection. There is a standard definition of correlation and there are some alternative measures of association to choose from, but they are all intended to measure the degree to which one variable "goes with" another.

---

8. The problem then becomes what to do with the identified relevant features in constructing a predictive model; that is, more a task of model selection and will be discussed later. See infra Part II (discussing different approaches to model selection).

9. See R.K. Templeton & J. Franklin, Adaptive Information and Animal Behaviour, 10 Evolutionary Theory 145, 145–46 (1992) (discussing traffic light example and adaptive information).

Features may be irrelevant to prediction in two ways: A feature is either not correlated with the variable being predicted, or it is correlated but is redundant because other features provide the same information, as they are highly correlated with it. Work in data mining concentrates on finding a suitable small subset of features, all relevant and, as far as possible, not redundant, which competently predict the target. There are some subtleties about the relevance of sets of features (as opposed to individual features), since two features could be relevant in combination although they are not relevant individually.[10]

Knowledge of the relevance of features can come in two ways. Either one measures correlation in the data, or one brings to bear prior knowledge of relevance—or perhaps more often, irrelevance. In *Shonubi*, discussed in Cheng and earlier, where the amount of drugs smuggled by Shonubi had to be estimated from data on "similar" drug smugglers, there was a reasonable prior belief that being a Nigerian drug mule was relevant to the amount of drugs smuggled, while having a day job as a toll collector was not.[11] Humans' prior beliefs on frequencies—e.g., whether eating colored mushrooms is often followed by illness, whether rocks typically have lions behind them—have been much studied and often found to be sound. But they have some persistent biases, such as a tendency to overweight very rare disasters such as air crashes. Therefore, any alleged prior knowledge is an important matter of evidential weight and thus important to reach a correct decision. So it should be subject to scrutiny in the usual way, not by statistical formulas, but by such means as a committee of experts or cross-examination in court.[12]

Having selected the features relevant to the prediction task, one then wishes to create a model. That is, one decides on the mathematical form of the relationship between the relevant features and the target. In the classic reference class problem, the outcome is predicted by a very simple function of the statistics of the reference class. For example, in *Shonubi*, the estimate of drugs smuggled by Shonubi was the average of drugs smuggled by members of the reference class: Nigerian drug mules at JFK during the time period. The question then is, how should the reference class be chosen, once the set of relevant—or strongly relevant—features have been identified?

There is a unique natural choice: The correct reference class is that defined by the intersection of the relevant features. If being Nigerian, being a drug mule, being at JFK, and being in the time period are all

---

10. Hopefully such problems are rare. There are a number of algorithms available for searching a database and finding sets of relevant features.

11. See Cheng, supra note 2, at 2082 (discussing statistical comparison in United States v. Shonubi, 895 F. Supp. 460, 466 (E.D.N.Y. 1995)).

12. See, e.g., James Franklin et al., Evaluating Extreme Risks in Invasion Ecology: Learning from Banking Compliance, 14 Diversity & Distributions 581, 582 (2008) (discussing how expert committees "encourage[d] care and transparency" in Australian import biosecurity agency analysis).

reasonably believed to be relevant to the amount of drugs smuggled, and there is no evidence that any other feature on which data is available is relevant, then the ideal choice of reference class is Nigerian drug smugglers at JFK in the time period.

The reference class problem then divides into two, depending on whether this ideal choice of reference class is usable. It is usable if there is a sufficiently large number of cases in it for a reliable estimate of the target. A data set that is too small, or even empty, will not support reliable estimates, since there is too much chance involved in which few cases happened to land in the set.[13] To know whether the data is enough to ensure reliability of the estimate, one consults standard statistical theory on the variance or standard deviation of the estimate in question. That is, one asks how variable the estimate is given the sample size: the smaller the sample, the more variable and thus unreliable the estimate. For example, if n drug smugglers are found with amounts $x_1, \ldots, x_n$ of drugs, of which the average is $\bar{x}$, then the standard deviation of $\bar{x}$ is approximately the standard deviation of the original n observations divided by $\sqrt{n}$. Or, if the problem is to estimate a probability based on a proportion in a small reference class—for example, eighteen of twenty cars on the surveillance video went through a red light, so the chance the defendant's car went through a red light is ninety percent—then the reliability of the estimate is given by calculating a "confidence interval."[14] This calculation again improves with the square root of the number of instances. Thus, we can quantify how much increasing the size of the reference class increases the reliability of the estimate, and how unreliable a very small reference class is.

If, on the other hand, the reference class defined by the intersection of relevant features is too small for reliable inference, or perhaps even empty, one is still left with useful information in the wider classes defined by taking some but not all of the relevant features. The statistics in those different classes will usually be different. For example, Nigerian drug smugglers in general and drug smugglers at JFK may have different averages of drugs smuggled, even though one may not have data on Nigerian drug smugglers at JFK. The problem then is how to combine the statistics in the different classes in which an individual lies, in order to make an estimate applicable to the individual case. A paradigm of the

13. Reichenbach said when coining the phrase "reference class problem" that it should be "the narrowest class for which reliable statistics can be compiled," which is correct, except that one does not narrow a relevant reference class by splitting it according to irrelevant attributes. Hans Reichenbach, The Theory of Probability 374 (1949).

14. See generally Lawrence D. Brown, T. Tony Cai & Anirban DasGupta, Interval Estimation for a Binomial Proportion, 16 Stat. Sci. 101, 101 (2001) (discussing "interval estimation of the probability of success in a binomial distribution"). Calculators that measure the confidence interval are available online. See Binomial Proportion Confidence Interval: Web-based Calculators, Wikipedia, at http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval (last visited Feb. 12, 2010) (on file with the *Columbia Law Review*) (providing links to web-based calculators).

problem is:

> Suppose we have under observation a certain Jones, who is found to be a Texan and a philosopher. We know that 99 per cent of all Texans are millionaires, and that only 1 per cent of all philosophers are millionaires (and we have no information about the class of Texan philosophers). On that evidence, what should we conclude about whether Jones is a millionaire? We would know the probability of Jones being a millionaire, given either one of those pieces of information, but what should be conclude when we have both?[15]

This problem unfortunately is unsolved. If the estimates coming from the different classes are close to one another, then of course it does not matter much which one is used, as the estimates in effect concur. If, as in the example, the estimates conflict, then any combination of them, even if correct, is unreliable. In that case, the issue may come down to whether one estimate is better on intuitive or other external grounds. For example, one may be reduced to arguing in court whether being Texan or being a philosopher is known to be more relevant to wealth.

We come now to the choice and fitting of models, where it is determined how the data deemed relevant will be used to make the estimate.

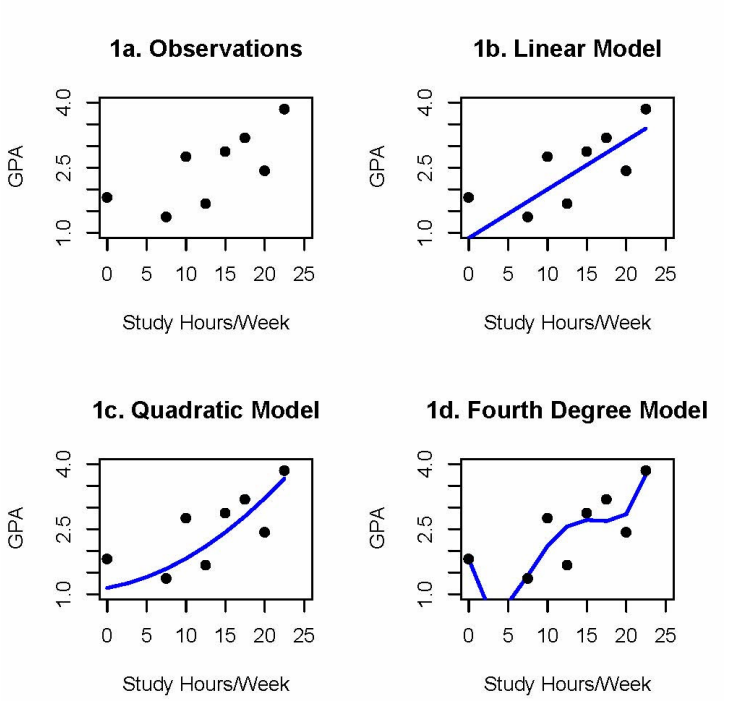
II. MODELS: SIMPLE OR SMOOTH?

Cheng poses the reference class problem as one of "model selection."[16] Statistical models comprise a range of techniques for deciding the form of the method to be applied to data before the parameters of the model are chosen by fitting it to the data. For example, one may decide that a linear, line of best fit model is applicable, and then find the slope of the line by fitting it to the data. That is illustrated in the top two panels of Cheng's Figure 1, which show a typical dataset of observations (Figure 1a) and a straight line of best fit to the data (Figure 1b).[17] The line expresses the best linear relationship between the two measured quantities—here, study hours and GPA—allowing a prediction of GPA from study hours for data not yet observed.

It is possible to fit more complex models, with quadratic, fourth degree or other formulas, as illustrated in Figures 1c and 1d, in the hope of obtaining more accurate predictions. That hope may or may not be realized. Sometimes more complex models appear to do better than a simple straight line (as in Figure 1c), while sometimes they appear to have become overcomplicated and do worse (as in Figure 1d). Cheng recommends choosing simple models—that is, ones with few parameters—with the degree of complexity chosen according to Akaike's

15. S.F. Barker, Induction and Hypothesis 76 (1957).
16. Cheng, supra note 2, at 2090–92.
17. Id. at 2091.

Information Criterion or some similar formula.[18] That is a respectable option, according to the statistical literature, and gives reasonable results when it comes to prediction. But it is not the only option, and it is arguably not well grounded in theory. An alternative to that "Ockhamist" approach, as it might be called in view of its emphasis on simplicity, is a "smoothing" approach, which does not care about the complexity of models as long as they are smooth, that is, varying little from point to point.



Cheng Fig. 1: Example Fits to Observed Datapoints

To explain the contrast between the two styles of models, let us go back to Cheng's example of fitting a curve to noisy one-input, one-output data.[19] The problem is: Given some points generated by an underlying but unknown function, perhaps with noise, how can you fit the best curve to them? In this sense, "best" would mean the ability to predict new points that would be generated by the same process.

One good feature of the problem is that, on the whole, the eye is quite competent to judge fit. One can see that the line in Figure 1a is too straight, the curve in Figure 1d too irregular, and the smooth curve in Figure 1c is about right. The eye, of course, does not see the formulas

18. Id. at 2093–94.
19. Id. at Part II.A (detailing model selection problem).

or the number of parameters in them, but only the smoothness of the fit. Another positive feature of the problem is that there are several mathematical methods available, implemented in software, that give more or less the same answer—polynomial regression, neural networks, smoothed splines, and kernel smoothing, among others. (To say that they give the same answer does not mean that they give the same or similar formulas, but similar graphs of the fitted curve.) The disadvantage of the problem is that there is still no agreement on the basics of the theory, the cause of the agreement between methods, or what feature the answer has that makes it the right answer. There are, in fact, two radically different views on the theory: an Ockhamist view and an anti-Ockhamist view.

According to the view based on Ockham's razor, it is a matter of choosing the right number of parameters for the curve. If one decides to fit a polynomial, one can choose a line, a quadratic, a cubic, and so on. A line is described by two parameters, intercept and slope, that are to be chosen by, or "fitted to," the data. A quadratic needs three parameters, a cubic four, and so on. If one chooses too few parameters—a too simple curve—it fails to fit the data well. On the other hand, if one chooses too many, the curve easily fits the actual data, but because it wobbles like jelly in response to the details of the actual data set, it would have been different if fitted to another data set generated by the same process. Hence it predicts poorly: It is said to have "overfitt[ed]" the data, to be "fitt[ed] to the noise," or to have "high variance."[20]

If one decides to fit one of a fixed family of parametric curves, such as polynomials, there is a reasonably well-established theory on how to choose the right number of parameters for a particular data set, and on why that number is correct. Bayesian statisticians with a background in physics have provided an analysis based on the precision of a prior distribution with few parameters versus the imprecision of one with many: A model with many trainable parameters "hedges its bets," so to speak, by being ready to fit anything, and hence is less "falsifiable" than a more precise one.[21] One speaks of "O[ckham's] hill," which has a peak at the correct number of parameters.[22] Statisticians who are not card-carrying Bayesians have a similar theory, which issues in a prescription called the Akaike Information Criterion for the number of parameters.[23]

---

20. See, e.g., Wallace E. Larimore & Raman K. Mehra, The Problem of Overfitting Data, 10 Byte 167, 168 (1985) ("[O]verfitting lessens the predictive value of the model.").

21. See Prasanta S. Bandyopadhayay, Robert J. Boik & Prasun Basu, The Curve-Fitting Problem: A Bayesian Approach, 63 Phil. Sci. S264 (1996) (using Bayesian theorem to solve curve-fitting problem); William H. Jeffreys & James O. Berger, Ockham's Razor and Bayesian Analysis, 80 Am. Sci. 64, 68 (1992).

22. See David J.C. MacKay, Bayesian Interpolation, 4 Neural Computation 415 (1992) (arguing Bayesian analysis infers values regularizing constants and noise levels, leading to effective number of parameters determined by data set).

23. See Malcolm Forster & Elliot Sober, How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions, 45 Brit. J. for Phil. Sci. 1

That works well and is reasonably convincing in the context of physics, where there is a prior expectation that simple models can often be found. It is not so clear that, where things are expected to be complex, as in economic modeling, taking a simple model is the only or best method of producing a falsifiable model, that is, one that will be found to perform well on new data.

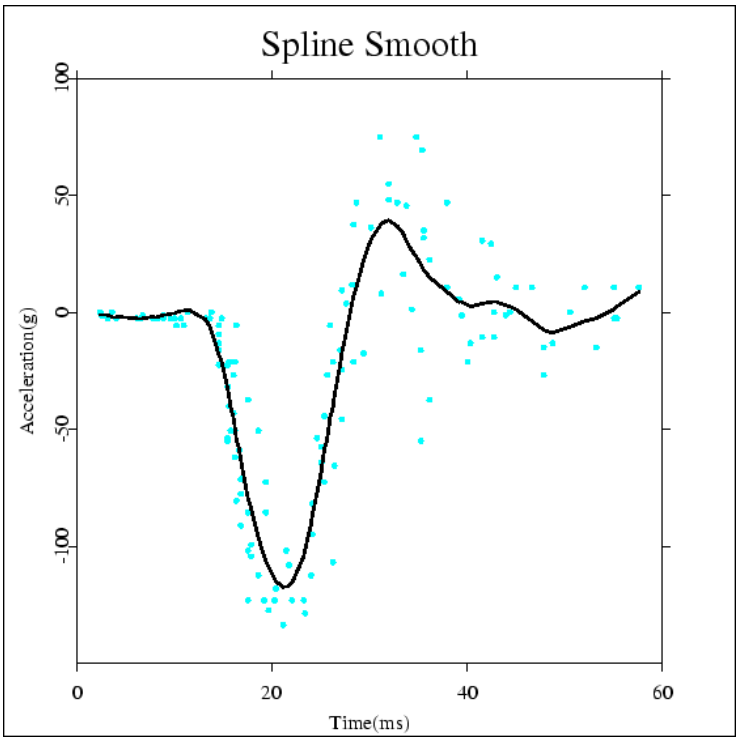


Fig. 2: Spline Smoothing of Observed Datapoints[24]

In response, the anti-Ockhamist school makes two points. Firstly, "having few parameters" is not the same thing as "simple": If one takes any wobbly curve, such as the cross-section of a piece of corrugated iron, it can stand in place of the straight line in the above example and generate a family of curves. It has not been explained why the bottom level of the family should be, for example, straight.[25] More importantly, some other methods which give much the same answer graphically, such as smoothed splines, achieve the result by finding a curve that is

(1994); I.A. Kieseppä, Akaike Information Criterion, Curve-Fitting, and the Philosophical Problem of Simplicity, 48 Brit. J. for Phil. Sci. 21 (1997).

24. B.W. Silverman, Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting, 47 J. Royal Stat. Soc'y B. 1, 9 (1985).

25. Kieseppä, supra note 23 (analyzing solutions to problems of "bumpier curves" and "smoother curves"); André Kukla, Forster and Sober on the Curve-Fitting Problem, 46 Brit. J. for Phil. Sci. 248, 249 (1995) (same).

sufficiently smooth, but not too smooth—but the curve does not have few parameters. Smoothing methods work by replacing the datapoints with averages of nearby datapoints: The value of the curve at any point is the weighted average of nearby datapoints. Different methods are distinguished by different technical decisions on how to weigh and determining how close is "nearby."

Given a smoothing method, the main problem is to decide on the right degree of smoothing: smooth too much and there is just a straight line that has lost most of the structure of the data, smooth too little and the result wobbles around fitting the idiosyncracies of the individual data set. In most circumstances, the right degree of smoothing is determined by the method of cross-validation, which calculates how well the estimate would predict a datapoint if it were left out: If the curve would be little changed by leaving out a datapoint, and would also predict that datapoint well, the degree of smoothing is correct.[26]

Further, in some related contexts, although simplicity does lead to good results, one can see that complexity is even better. Studies of machine learning, which is in principle a higher-dimensional version of curve-fitting, show that complicating the result, while preserving its behavior on the old data, can improve its performance.[27] Leo Breiman goes so far as to speak of "two cultures" in statistics: an older style that looks for simple and explanatory models of data, and a more contemporary style that embraces large and internally complex "black-box" predictors such as neural nets and random forests, as long as they are equipped with methods of smoothing to prevent overfitting of the data.[28]

The natural conclusion to reach is that simplicity, or fewness of parameters, is not in itself desirable in curve fitting and related contexts, but only works because it is normally used in such a way as to correlate with smoothness, which is what really enables prediction. Prediction is what counts, and simplicity may or may not help it.

CONCLUSION

Statistical methods do have advice to offer on how courts should judge quantitative evidence, but in a way that supplements normal intuitive legal argumentation rather than replacing it by a formula. In cases like *Shonubi* and those involving real estate valuation or the

---

26. See Grace Wahba, Spline Models for Observational Data 45–49 (1990); Richard R. Picard & R. Dennis Cook, Cross-Validation of Regression Models, 79 J. Am. Stat. Ass'n 575 (1984).

27. See Pedro Domingos, The Role of Occam's Razor in Knowledge Discovery, 3 Data Mining and Knowledge Discovery 409 (1999); C. Schaffer, Overfitting Avoidance as Bias, 10 Machine Learning 153 (1993); G.I. Webb, Further Experimental Evidence Against the Utility of Occam's Razor, 4 J. Artificial Intelligence Res. 397 (1996).

28. Leo Breiman, Statistical Modeling: The Two Cultures, 16 Stat. Sci. 199, 221–22 (2001) (advocating active monitoring to protect against overfitting).

measurement of environmental risks, there is relevant quantitative data and hence a need for technical statistical advice. It is crucial to know what properties of the data are statistically relevant to the inferential task, the answer to which determines the appropriate reference class in which to take counts. Knowledge of the relevance of properties is of two kinds. The first, commonsense or scientific knowledge of causes and symptoms, is subject to the usual style of intuitive reasoning and challenge in the courtroom. The second, obtained from statistical methods such as variable selection and the fitting of models, is also crucial when there is significant reliance on inference from a set of quantitative data. The statistical techniques should neither dominate nor be dominated by intuitive reasoning. Lawyers need to understand both styles of reasoning in order to integrate them.

Preferred Citation: James Franklin, *Feature Selection Methods for Solving the Reference Class Problem: Comment on Edward K. Cheng, "A Practical Solution to the Reference Class Problem*," 110 COLUM. L. REV. SIDEBAR 12 (2010), http://www.columbialawreview.org/sidebar/volume/110/12_franklin.pdf.